

# Statistical information extraction : a tutorial and real-life case study

Marc Vilain  
The MITRE Corporation

Presented at ASIAS symposium  
July 27-28, 2009



# What is information extraction?

- **Identification of entities, relationships, and events in language sources**
  - People, places, organizations, vehicles, ...
  - Employment, family relations, domicile, ...
  - Merger/acquisition, kidnapping, disease outbreak, ...
- **Basis for indexing, mapping, database creation, statistical analysis, knowledge discovery ...**
- **Application areas**
  - Law enforcement and national security
  - Business intelligence (growth area)
  - Epidemiology mapping

# ASAP reports: entities

airport

runway

In **MKG** we were given clearance of 'taxi to **runway 14** via **36**.' ...  
While taxiing to **14** we were told to hold short for landing ...  
We were confused as to which line was the **hold short line for runway 14** as there were **two hold short lines** in close proximity. ...

We unintentionally crossed the **hold short line for runway 14** thinking it was the **hold short for runways 18/36** ...

The event occurred due to confusion of the **runway markings**, airport map page, and information [from ATC] ...

The **markings for the hold short** and **markings for runway 18/36** did not look right ...

I was also preoccupied with running checklists and informing the flight attendant of our immediate departure ...

ground markings



**MITRE**

© 2009, The MITRE Corporation

# ASAP reports: events

airplane motion

In MKG we were given clearance of '**taxi to runway 14 via 36.**' ...  
While **taxiing to 14** we were told to **hold short for landing** ...  
We were confused as to which line was the hold short line  
for runway 14 as there were two hold short lines  
in close proximity. ...

incident

We unintentionally **crossed the hold short line for runway 14**  
thinking it was the hold short for runways 18/36 ...

The event occurred due to confusion of the runway markings,  
airport map page, and information [from ATC] ...

procedure

The markings for the hold short and markings for runway  
18/36 did not look right ...

I was also preoccupied with **running checklists** and **informing  
the flight attendants** of our immediate departure ...

# ASAP reports: sentence classification

confusion

In MKG we were given clearance of '**taxi to runway 14 via 36.**' ...  
While **taxiing to 14** we were told to **hold short for landing** ...

We were confused as to which line was the **hold short line for runway 14** as there were **two hold short lines** in close proximity. ...

responsibility

We unintentionally **crossed the hold short line for runway 14** thinking it was the **hold short for runways 18/36** ...

The **event occurred due to confusion** of the **runway markings**, airport map page, and information [from ATC] ...

The **markings for the hold short** and **markings for runway 18/36** did not look right ...

I was also preoccupied with **running checklists** and **informing the flight attendants** of our immediate departure ...

attention

# Case study: Business news

Callisto markup tool

The screenshot shows a window titled "Callisto - all-02-28-2007.xml.crf.scorable" with a menu bar (File, Edit, Format, Tools, Help). The main text area contains a news article snippet with several entities highlighted in blue. Below the text is a table of extracted entities with columns for "Text" and "Type".

USFAB01282320070228  
Mylan interested in Merck KGaA drug unit-sources  
Wed Feb 28, 2007 8:35AM EST

LONDON/FRANKFURT, Feb 28 (Reuters) - U.S.-based Mylan Laboratories Inc (MYL.N: Quote, Profile, Research) is interested in bidding for Merck KGaA's (MRCG.DE: Quote, Profile, Research) generics business, joining a growing of band of firms lining up offers, people familiar with the situation said on Wednesday.

First-round offers for the unit, which analysts expect to sell for more than 4 billion euros (\$5.29 billion), are due by the middle of March and Merck is hoping to close the sale by May or June, the sources added.

ENAMEX TIMEX NUMEX

Filter on Type <ALL>

Text	Type
Mylan	ORGANIZATION
Merck KGaA	ORGANIZATION
LONDON	LOCATION
FRANKFURT	LOCATION
Reuters	ORGANIZATION
U.S.	LOCATION
Mylan Laboratories Inc	ORGANIZATION
Merck KGaA	ORGANIZATION
First-round	ORGANIZATION

Font: 16pt. Default | Charset: UTF-8 | Task: MUC+Fin Task

Entities: money, companies, ...

Events: mergers and acquisitions

# Approaches to extraction

## ▪ Rule-based

- Entities extracted by pattern-matching rules
- MR. XXX → MR. <PERSON>XXX</PERSON>
- Pattern rules must be built and ordered by hand
- To port to new task, rule base must be manually modified; for rule-based tools, this usually requires hiring the vendor

## ▪ Statistical

- Entities extracted by statistical evidence-weighting
- $P(w_0 \approx \text{person} \mid w_{-1} = \text{"MR."})$
- Probabilities are estimated from training data
- To port to new task, retrain estimates with task-specific training data

Probability that current word is person-typed, given that previous word is "MR."

# Statistical extraction: Business news how-to

**(0) Train baseline system from legacy newswire data**

**(1) Manually annotate M&A data**

Tag 150 news stories (one day of news) to set guidelines

Tag 3 additional news days for training/eval

**(2) Create additional training data**

Manually tag related news (hot stocks, general biz, ...)

Partially auto-tag more M&A news

**(3) Train statistical CARAFE models**

Implement various tricks to handle auto-tagged data and capture discourse effects

Evaluate performance

# Technical tricks

- **Resources**

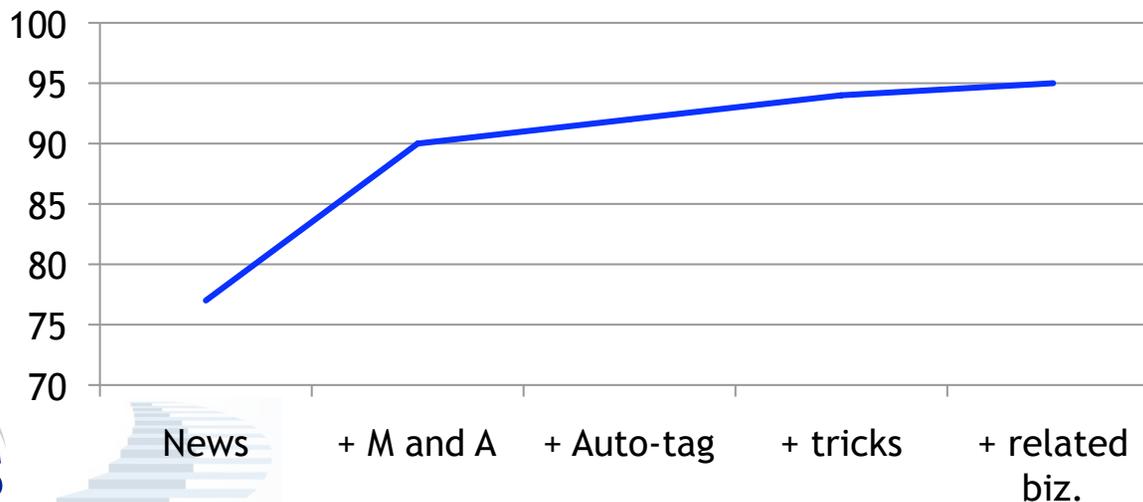
- Gazetteers capture states/provinces, known first names
- Other lists capture days of week, months, etc.

- **Long-distance dependencies**

- “Thomas White ... . White was named global head of ...”
- First instance of “White” predicted from context (“Thomas”), but second instance lacks predictive context
- Use feature-copying trick:  
 $\Phi(w): w'_{-1} \in \textit{first-names}$  for some  $w'=w$  in the narrative
- Also effective for banks (Global Credit Union ... Global) and companies (Zelcor Widgets Corp. ... Zelcor)

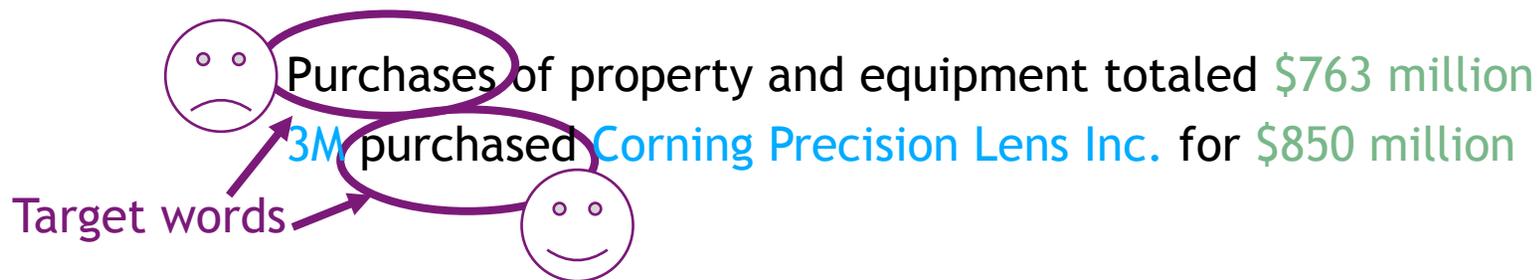
# Evaluation

- Overall performance: F=95 (R=95, P=95)
- Performance for specific entities
  - Organization F=94 (R=94, P=95)
  - Person F=92 (R=92, P=92)
  - Location F=93 (R=92, P=94)
  - Date F=97 (R=98, P=97)
  - Money F=98 (R=97, P=99)



# Statistical identification of events

- Multi-stage process based on statistical classifiers
- Step 1: classify target words as to whether they're event-denoting (maximum entropy models)



- Step 2: assign phrases to event roles (more classifiers)

3M purchased Corning Precision Lens Inc. for \$850 million

Buying company

Acquired company

Amount

# Conclusions

- **Statistical methods just simply work**
  - Trainable methods allow for rapid adaptation to new tasks
  - Great success with many tasks and many sources
  - Frequently much better performance than *un-adapted* system (most out-of-the-box non-trainable tools)
  - Engineering requirements are relatively few
  - Manual annotation burden is relatively slight
- **Missing: statistical tools that are end user-trainable**
  - Annotation and tool configuration remain specialized, though teachable, skills